Robust Speech Recognition Based on a Bayesian Prediction Approach

Hui Jiang, Keikichi Hirose, *Member, IEEE,* and Qiang Huo, *Member, IEEE*

***Abstract*—In this paper, we study a category of robust speech recognition problem in which mismatches exist between training and testing conditions, and no accurate knowledge of the mismatch mechanism is available. The only available information is the test data along with a set of pretrained Gaussian mixture continuous density hidden Markov models (CDHMM's). We investigate the problem from the viewpoint of Bayesian prediction. A simple prior distribution, namely constrained uniform distribution, is adopted to characterize the uncertainty of the mean vectors of the CDHMM's. Two methods, namely a model compensation technique based on Bayesian predictive density and a robust decision strategy called Viterbi *Bayesian predictive classification* are studied. The proposed methods are compared with the conventional Viterbi decoding algorithm in speaker-independent recognition experiments on isolated digits and TI connected digit strings (TIDIGITS), where the mismatches between training and testing conditions are caused by: 1) additive Gaussian white noise, 2) each of 25 types of actual additive ambient noises, and 3) gender difference. The experimental results show that the adopted prior distribution and the proposed techniques help to improve the performance robustness under the examined mismatch conditions.**

***Index Terms*—Bayesian predictive classification, minimax decision, plug-in maximum *a posteriori* decision, predictive density, Viterbi Bayesian predictive classification.**

I. Introduction

In THE past decade, tremendous advances have been achieved in automatic speech recognition (ASR) (e.g., see [21] for a sample of the state-of-the-art). These advances promise to make speech recognition technology readily available to the general public. However, as speech recognition systems are applied in real-world applications, they must be operated in situations where it is not possible to control the acoustic environment and application conditions. This may result in a serious mismatch between the training and testing conditions, which often brings about such a drastic degradation in performance that these systems usually fail in the real-field applications.

A substantial amount of work has been performed in robust ASR area to achieve performance robustness under various types of mismatches such as different kinds of additive ambient noises; convolutional channel/transducer mismatch; acoustic variations caused by inter- and/or intraspeaker variability, different accents, stress, different speaking styles, specific limitations in various tasks, etc. (see reviews in, e.g., [3], [6], [20], and [22]). Among many promising approaches, one is the feature (e.g., [1] and [29]) and/or model (e.g., [23] and [29]) compensation techniques to remove or reduce the acoustic mismatches between the test data and a given set of speech models. For this type of approach, some prior knowledge about the mechanism of mismatches is necessary to design a suitable form of mapping function. Then the *nuisance parameters* of the mapping function can be estimated based on a certain criterion such as maximum likelihood (ML) or maximum *a posteriori* (MAP) only with small amount of adaptation data or test data themselves.

However, in practice we generally have no idea about the sources of variability in speech signals, and no full knowledge to figure out the mechanism of mismatches between training data in the laboratory and test data in real field. In the extreme case, the only available information is the test data along with a set of pretrained speech models. An attractive approach that does not need accurate knowledge of the mismatch mechanism and adapts the speech models using only the test data is the so-called online Bayesian adaptive learning algorithm (e.g., [9]–[11]). This approach is suitable for those applications involving a recognition session which consists of a number of testing utterances. Besides, some recent approaches have focused on modifying the decision rule and the decision parameters so that part of the mismatch can be compensated and the decision performance can be improved. This scheme becomes a potential approach for robust speech recognition because it need not make rigid assumptions about sources of distortion. One such approach is the *minimax classification* method [25], which assumes the best decision parameters for the given test data lie in the neighborhoods of the given parameters and adjusts the decision rule and the corresponding parameters accordingly. The minimax classification is thus geared to protect against the possibility of the worst mismatch. The main disadvantage of the minimax approach is that it usually does not perform nearly as well as those techniques that use some prior information of the possible mismatches. Another disadvantage is that it can not be easily extended

Manuscript received July 28, 1997; revised October 14, 1998. The associate edtior coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

H. Jiang was with the Department of Information and Communication Engineering, University of Tokyo, Tokyo 113-8656, Japan. He is now with the Department of Electrical and Computer Engineering, University of Waterloo, Ont. N2L 3G1, Canada (e-mail: hjiang@crg3.uwaterloo.ca).

K. Hirose is with the Department of Information and Communication Engineering, University of Tokyo, Tokyo 113-8656, Japan (e-mail: hirose@gavo.t.u-tokyo.ac.jp).

Q. Huo was with the ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. He is currently with the Department of Computer Science and Information Systems, University of Hong Kong, Hong Kong (e-mail: qhuo@cs.hku.hk).

Publisher Item Identifier S 1063-6676(99)04631-3.

to perform continuous speech recognition (CSR) because the combination of uncertainty neighborhoods surrounding the model parameters that need to be examined can become quite large [25], [28].

In this paper, in the viewpoint of Bayesian prediction, we investigate two techniques to address the robust recognition problem in the above mentioned context to mitigate to some extent the difficulties of the minimax approach. We model each speech unit with a Gaussian mixture continuous density hidden Markov model (CDHMM). In the first technique, we assume some uncertainty of the CDHMM parameters and use the *Bayesian predictive density* of each Gaussian mixture component to serve as the compensated distribution of this component [15]. We thus call it *Bayesian predictive density based model compensation* (BP-MC) method. In this method, the decoding algorithm for speech recognition still uses the conventional *plug-in MAP* decision rule (see the discussions in, e.g., [2], [12], and [13]). In the second technique, by modifying directly the plug-in MAP decision rule, we have recently adopted a new robust decision strategy called *Bayesian predictive classification* (BPC) approach [12], [13], [16], [27] for robust speech recognition. We present here an approximate BPC algorithm called *Viterbi Bayesian predictive classification* (VBPC) [16]. We gather together and summarize in this paper those results scattered in [15]–[19] and some new experimental results as well, in order to make it more accessible to the general readership. Whenever possible, we use the same notations as those in [10]–[13].

The remainder of the paper is organized as follows. At first, several basic decision rules available for ASR are briefly introduced in Section II. Next, we describe the proposed techniques, namely BP-MC and VBPC approaches in Sections III and IV, respectively. To examine the viability of the above proposed approaches, a series of comparative experiments are conducted on two speech databases: ATR isolated Japanese digit database and TIDIGITS English connected digit string database. The corresponding experimental results are reported and brief discussions are presented in Section V. Finally, our conclusions are summarized in Section VI.

II. Decision Rules for Automatic Speech Recognition

In order to clarify the motivations of our work and to facilitate the discussions in the succeeding sections, we derive and repeat here some of the discussions originally presented in [12] and [13]. Let's view a *word* W and the associated acoustic observation $\mathbf{X}$ (usually, a feature vector sequence) as a jointly distributed random pair $(W, \mathbf{X})$. Depending on the problem of interest, the meaning of the *word* here could be any linguistic unit, e.g., a phoneme, a syllable, a word, a phrase, etc. Also note that in this paper we simply use the same symbol to denote both the random variable and the value it may assume. Suppose the *true* joint distribution of $(W, \mathbf{X})$ could be modeled by a *true parametric family* of probability density function (pdf) as $p(W, \mathbf{X}) = p_\Lambda(\mathbf{X}|W) \cdot p_\Gamma(W)$, where $p_\Lambda(\mathbf{X}|W)$ is known as acoustic model with parameters Λ and $p_\Gamma(W)$ as language model with parameters Γ. Further, suppose that we have the full knowledge on the parameters (Λ, Γ) of the above distributions. Then, the optimal decoder (speech recognizer) which achieves *expected* minimum *word* recognition error rate is the following MAP decoder:

$$\begin{aligned}\hat{W} &= \arg\max_W p(W|\mathbf{X}) \\ &= \arg\max_W p(W, \mathbf{X}) \\ &= \arg\max_W p_\Lambda(\mathbf{X}|W) \cdot p_\Gamma(W) \end{aligned} \tag{1}$$

where $\mathbf{X}$ is the observation and $\hat{W}$ is the recognition result. The decision rule in (1) is generally referred to as *optimal MAP decision rule*.

A. Plug-In MAP Rule

However, in practice, neither do we know the *true* parametric form of $p(W, \mathbf{X})$, nor its *true* parameters. Therefore, the above optimal speech recognizer will never be achievable; we can only approximate it. A simple heuristic solution is first to assume a parametric form for $p(W, \mathbf{X})$ and then to estimate its parameters from training data using a parameter estimation technique (e.g., ML, MAP, discriminative training, etc.). Then, we *plug in* the estimate $(\tilde{\Lambda}, \tilde{\Gamma})$ to the optimal but unavailable rule in (1) in place of the correct but unknown (Λ, Γ) to obtain a *plug-in MAP rule*. The plug-in MAP rule has been widely adopted by the current speech recognizers. The performance of plug-in MAP decision rule depends on the choice of estimation methods, the nature and size of the training data, and the degree of the mismatch between training and testing conditions.

B. Minimax Rule

As mentioned above, we generally have no full knowledge to figure out the *true* parameters of models or/and decision rule. Instead of using the estimated values as in the plug-in MAP rule, we assume that the unknown *true* parameters Λ are uncertain (random variables) and randomly distributed in a *neighborhood* region Ω around the estimated ones. If we have no further knowledge about Λ, a reasonable decision is to warrant the optimal outcome (e.g., minimum error) in the possibly worst-case condition (e.g., maximum mismatch) [7]. Such a *minimax decision rule* which minimizes the *upper bound* of the *worst-case probability of classification error* has been proposed in [25] as

$$\hat{W} = \arg\max_W \Big[\max_{\Lambda \epsilon \Omega} p(\mathbf{X}|\Lambda, W) \cdot p(W)\Big]. \tag{2}$$

Therefore, the minimax approach is considered to be the most conservative decision strategy.

C. Bayesian Predictive Classification Rule

An attractive compromise between the risky *plug in MAP rule* and the overdue conservative minimax approach is the decision strategy BPC [5], which can somehow make use of the prior knowledge (albeit crude) about the *possible* mismatch, and at the same time take into account its uncertainty to compensate accordingly for the possible severe mismatch. As in [4], [8], and [10], we use a prior pdf $p(\Lambda|\varphi)$

with hyperparameter φ to represent our knowledge about the uncertainty of the unknown parameters Λ. An *optimal Bayes solution* is to choose a speech recognizer which minimizes the *overall recognition error* when the average is taken both with respect to the sampling variation in the expected testing data and with respect to the uncertainty described by the prior pdf $p(\Lambda|\varphi)$. Such a BPC rule is as follows:

$$\begin{aligned}\hat{W} &= \arg\max_W \tilde{p}(W|\mathbf{X}) \\ &= \arg\max_W \tilde{p}(W, \mathbf{X}) \\ &= \arg\max_W \tilde{p}(\mathbf{X}|W)\cdot p(W)\end{aligned} \tag{3}$$

where

$$\tilde{p}(\mathbf{X}|W) = \int p(\mathbf{X}|\Lambda, W)p(\Lambda|\varphi, W)\,d\Lambda \tag{4}$$

is called the predictive pdf of the observation $\mathbf{X}$ given the word W. Generally speaking, the computation of this predictive pdf is the most difficult part of the BPC approach. Not like other approaches such as the *plug-in MAP* and the minimax where only a single set of values (called point estimate [8], e.g., mode, mean, etc.) of prior distribution is taken into account, as shown in (4), the whole function of prior distribution can be considered for decision-making in the Bayesian prediction procedure.

III. Bayesian Predictive Density Based Model Compensation Approach

There are many possible ways to apply Bayesian prediction to CDHMM-based speech recognition. A straightforward approach is described in this section. In this approach, instead of directly modifying the basic decision rule, we assume the CDHMM parameters are uncertain. Then we use the *Bayesian predictive density* of each Gaussian mixture component to serve as the compensated distribution of that component and plug these compensated distributions into the MAP decision rule in (1). We thus call the approach Bayesian predictive density based model compensation method, or shortly BP-MC method thereafter, to differentiate it from the BPC rule defined in (3) [12], [13].

We model each speech unit with an N-state CDHMM with parameter vector $\Lambda = (\pi, A, \theta)$, where π is the initial state distribution, $A = \{a_{ij}|1 \le i, j \le N\}$ is the transition matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\ldots,K}$ for each state i, where K denotes the number of Gaussian mixture in each state. The state observation pdf is assumed to be a mixture of multivariate Gaussian pdf's:

$$p(\mathbf{x}|\theta_i) = \sum_{k=1}^{K} \omega_{ik} f(\mathbf{x}|\theta_{ik}) = \sum_{k=1}^{K} \omega_{ik}\mathcal{N}(\mathbf{x}|m_{ik}, r_{ik}) \tag{5}$$

where the mixture coefficients ω_{ik}'s satisfy the constraint $\sum_{k=1}^{K}\omega_{ik} = 1$, and $\mathcal{N}(\mathbf{x}|m_{ik}, r_{ik})$ is the kth normal mixture component denoted by

$$\mathcal{N}(\mathbf{x}|m_{ik}, r_{ik}) \propto |r_{ik}|^{1/2}\exp\left[-\tfrac{1}{2}(\mathbf{x}-m_{ik})^t\, r_{ik}\,(\mathbf{x}-m_{ik})\right] \tag{6}$$

with m_{ik} being the D-dimensional mean vector and r_{ik} being the $D\times D$ precision (inverse covariance) matrix.

The BP-MC method adopted in the study can be simply described as follows.

1) For each mixture component $f(\mathbf{x}|\theta_{ik})$, a prior p.d.f. $p(\theta_{ik}|\varphi_{ik})$ with hyperparameters φ_{ik} is assumed to represent our knowledge about the uncertainty of the CDHMM parameters θ_{ik}.
2) A Bayesian predictive density is computed as

$$\tilde{f}_{ik}(\mathbf{x}) = \int f(\mathbf{x}|\theta_{ik})p(\theta_{ik}|\varphi_{ik})\,d\theta_{ik}. \tag{7}$$

3) Compute $\tilde{p}_i(\mathbf{x}) = \sum_{k=1}^{K}\omega_{ik}\tilde{f}_{ik}(\mathbf{x})$ and *plug* it into the decision rule in (1) in place of the state observation pdf $p(\mathbf{x}|\theta_i)$.

The choice of prior pdf $p(\theta_{ik}|\varphi_{ik})$ depends on the prior knowledge about both θ_{ik} and the mismatch in question. In this paper, as the first step, we only consider the uncertainty of the mean vectors of CDHMM with diagonal covariance matrices and assume they are uniformly distributed in a neighborhood of pretrained means. The similar uncertainty neighborhood of Λ as defined in [25] is adopted as follows:

$$\begin{aligned}\eta(\Lambda) = \{\Lambda\,|\,&\pi_i = \pi_i^*,\, a_{ij} = a_{ij}^*,\, \omega_{ik} = \omega_{ik}^*,\, r_{ik} = r_{ik}^*, \\ &|m_{ikd} - m_{ikd}^*| \le C\,d^{-1}\rho^d,\, 1 \le i \le N, \\ &1 \le k \le K,\, 1 \le d \le D\}\end{aligned} \tag{8}$$

where hyperparameters C ($C > 0$) and ρ ($0 \le \rho \le 1$) are used to control, respectively, the possible mismatch *size* and *shape*, and $\{\pi_i^*, a_{ij}^*, m_{ikd}^*, r_{ik}^*\}$ denote the pretrained model parameters. The constrained uniform distribution in the above uncertainty neighborhood is referred to as *less-informative* prior pdf to contrast with other more informative distributions (in terms of parametric form) such as the normal distribution.

We then have $\tilde{f}_{ik}(\mathbf{x}) = \prod_{d=1}^{D}\tilde{f}_{ikd}(x_d)$ with

$$\begin{aligned}\tilde{f}_{ikd}(x_d) &= \left(\frac{r_{ikd}^*}{2\pi}\right)^{(1/2)}\frac{1}{2Cd^{-1}\rho^d}\int_{m_{ikd}^*-Cd^{-1}\rho^d}^{m_{ikd}^*+Cd^{-1}\rho^d} \\ &\quad\cdot e^{-(1/2)r_{ikd}^*(x_d-m_{ikd})^2}\,dm_{ikd} \\ &= \frac{1}{2Cd^{-1}\rho^d}\left\{\Phi\left(\sqrt{r_{ikd}^*}\,(m_{ikd}^* - x_d + Cd^{-1}\rho^d)\right)\right. \\ &\quad\left.-\Phi\left(\sqrt{r_{ikd}^*}(m_{ikd}^* - x_d - Cd^{-1}\rho^d)\right)\right\}\end{aligned} \tag{9}$$

where

$$\Phi(y) = \frac{1}{\sqrt{2\pi}}\cdot\int_{-\infty}^{y} e^{-(x^2/2)}\,dx. \tag{10}$$

As a remark, in [30], a similar idea has been explored in the context of Bayesian speaker adaptation where a Gaussian prior pdf for mean vector is adopted.

IV. Viterbi Bayesian Predictive Classification Approach

In [12] and [13], we discuss how to apply the general BPC to CDHMM-based robust speech recognition, and finally

focus on an approximate BPC method called *quasi-Bayesian predictive classification* (QBPC). Here, we focus our study on another approximate BPC method, namely Viterbi BPC (VBPC) approach.

In the CDHMM case, due to the nature of the *missing data* problem in HMM formulation (see related discussions in [10], [12], and [13]), it is not easy to compute the true *predictive pdf*:

$$\tilde{p}(\mathbf{X}|W) = \sum_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W) p(\Lambda|\varphi, W)\, d\Lambda \quad (11)$$

where $\mathbf{s}$ is the unobserved state sequence and $\mathbf{l}$ is the associated sequence of the unobserved mixture component labels corresponding to the observation sequence $\mathbf{X}$. Consequently, some approximations are needed [12], [13]. One way to compute the approximate predictive pdf is to use the following Viterbi approximation:

$$\tilde{p}(\mathbf{X}|W) \approx \max_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W) p(\Lambda|\varphi, W)\, d\Lambda. \quad (12)$$

The resultant BPC rule is named as VBPC rule:

$$\hat{W} = \arg\max_{W} \left[p(W) \cdot \max_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W) \cdot p(\Lambda|\varphi, W)\, d\Lambda \right]. \quad (13)$$

As shown in (11), when the *missing data* $\{\mathbf{s}, \mathbf{l}\}$ is unknown, the summarization over all possible $\{\mathbf{s},\mathbf{l}\}$ makes the true predictive pdf unachievable. However, once $\{\mathbf{s},\mathbf{l}\}$ is given or hypothesized, the Bayesian prediction calculation becomes straightforward. Here we present a frame-synchronous Viterbi Bayesian search algorithm, which is extended from the conventional Viterbi search algorithm, to achieve the above VBPC rule.

- For every time instant, compute the predictive values for all active hypothesized partial paths, respectively.
- Then for each node in the network, merge all incoming partial paths via selecting the one with the largest predictive value.
- The selected path is propagated and its predictive value is recomputed according to the extended partial path.
- The above search procedure is repeated until the end of the utterance.

Given a test utterance $\mathbf{X} = (x_1, x_2, \cdots, x_T)$, CDHMM parameter vector Λ along with its prior pdf $p(\Lambda)$, the recursive search procedure for *approximately*[1] accomplishing the above VBPC rule (13) is described as follows.

1) Initialization: $t = 1$

$$\delta_1(i) = \tilde{\pi}_i \cdot \tilde{b}_i(x_1) \qquad 1 \le i \le N \quad (14)$$

$$\psi_1(i) = 0 \qquad 1 \le i \le N \quad (15)$$

where $\tilde{\pi}_i$ denotes the mean of the prior pdf of the HMM parameter π_i, i.e.,

$$\tilde{\pi}_i = \int \pi_i \cdot p(\Lambda)\, d\Lambda \quad (16)$$

and

$$\tilde{b}_i(x_t) = \int p(x_t|\theta_i) \cdot p(\Lambda)\, d\Lambda. \quad (17)$$

Here, $\delta_t(i)$ denotes the partial predictive value based on the optimal partial path arriving at state i at the time instant t. The corresponding best partial path is represented by a chain of points started from $\psi_t(i)$.

2) Recursion: For $2 \le t \le T$, $1 \le j \le N$, do

(2.1 Path-merging in state j, and update partial predictive value with respect to $\{a_{ij}\}$:

$$\overline{\delta}_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) \cdot \tilde{a}'_{ij}] \quad (18)$$

$$\psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) \cdot \tilde{a}'_{ij}]. \quad (19)$$

The $\tilde{a}'_{ij}$ is the mean of the posterior pdf of the a_{ij} based on the optimal partial path up to the time instant t, i.e.,

$$\tilde{a}'_{ij} = \begin{cases} \tilde{a}_{ij} & \text{for } i \ne j \\ \tilde{a}_{ij}^{(L_{ij})} \Big/ \tilde{a}_{ij}^{(L_{ij}-1)} & \text{for } i = j \end{cases} \quad (20)$$

where L_{ij} is the accumulated number of transitions from state i to state j based on the optimal partial path up to the time instant t; $\tilde{a}_{ij}$ denotes the mean of the prior pdf of the HMM parameter a_{ij}, and $\tilde{a}_{ij}^{(n)}$ correspondingly denotes the nth-order moment of a_{ij}, i.e.,

$$\tilde{a}_{ij} = \int a_{ij} \cdot p(\Lambda)\, d\Lambda \quad (21)$$

$$\tilde{a}_{ij}^{(n)} = \int a_{ij}^{n} \cdot p(\Lambda)\, d\Lambda. \quad (22)$$

(2.2 Update the partial predictive value with respect to state parameter θ_j:
If [it is the first time to involve state j in computation of $\delta_t(j)$],[2] **then**

$$\delta_t(j) = \overline{\delta}_t(j) \times \tilde{b}_j(x_t). \quad (23)$$

Else

$$\delta_t(j) = \overline{\delta}_t(j) \times \frac{\tilde{b}_j(x_{j_1}, x_{j_2}, \cdots, x_{j_{L_j}})}{\tilde{b}_j(x_{j_1}, x_{j_2}, \cdots, x_{j_{(L_j-1)}})} \quad (24)$$

where L_j is the accumulated number of feature vectors belonging to state j based on the optimal partial path up to the time instant t; x_{j_i} denotes the ith vector in the state j; and $\tilde{b}_j(x_{j_1}, x_{j_2}, \cdots, x_{j_{L_j}})$ denotes the contribution of data $\{x_{j_1}, x_{j_2}, \cdots, x_{j_{L_j}}\}$, residing at state j, to the partial predictive value $\delta_t(j)$:

$$\tilde{b}_j(x_{j_1}, x_{j_2}, \cdots, x_{j_n}) = \int p(x_{j_1}|\theta_j) \cdot p(x_{j_2}|\theta_j) \cdots p(x_{j_n}|\theta_j) \cdot p(\Lambda)\, d\Lambda. \quad (25)$$

[1] Strictly speaking, the search algorithm here is nonadmissible: It cannot completely warrant (13) in theory because the partial predictive value (i.e., δ_t) will possibly be recomputed partially in (24) during search.

[2] Including all states tied to state j.

3) Termination:

$$\tilde{p}(\mathbf{X}|W) \approx \max_i \delta_T(i) \tag{26}$$

$$s_T^* = \arg\max_i \delta_T(i). \tag{27}$$

4) Path (State Sequence) Backtracking: (4)

$$s_t^* = \psi_{t+1}(s_{t+1}^*) \qquad t = T-1, T-2, \cdots, 1. \tag{28}$$

In this section, we also only consider the uncertainty of the mean vectors of CDHMM with diagonal covariance matrices. So we have

$$\tilde{\pi}_i = \pi_i,\ \tilde{a}_{ij} = a_{ij} \quad \text{and} \quad \tilde{a}_{ij}^{(n)} = a_{ij}^n \quad \text{with} \quad 1 \le i, j \le N. \tag{29}$$

Moreover, we follow the same choice of the less-informative prior pdf $p(\Lambda)$ as in the last section [defined in (8), etc.] We then have

$$\begin{aligned}\tilde{b}_j(x_t) &= \sum_{l_t=1}^{K} \omega_{jl_t} \cdot \tilde{f}_{jl_t}(x_t) \\ &\approx \omega_{jl_t^*} \cdot \tilde{f}_{jl_t^*}(x_t) \\ &= \omega_{jl_t^*} \cdot \prod_{d=1}^{D} \tilde{f}_{jl_t^* d}(x_{td})\end{aligned} \tag{30}$$

where $\tilde{f}_{jl_t d}(x_{td})$ follows (9) and l_t^* denotes the mixture component label to which x_t is "closest," i.e.,

$$\begin{aligned}l_t^* &= \arg\max_{l_t} \left[\omega_{jl_t} \cdot \tilde{f}_{jl_t}(x_t)\right] \\ &= \arg\max_{l_t} \left[\omega_{jl_t} \cdot \prod_{d=1}^{D} \tilde{f}_{jl_t d}(x_{td})\right].\end{aligned} \tag{31}$$

Similarly, $\tilde{b}_j(x_{j_1}, x_{j_2}, \cdots, x_{j_n})$ is calculated based on the "closest" mixture component label sequence corresponding to the data $\{x_{j_1}, x_{j_2}, \cdots, x_{j_n}\}$:

$$\begin{aligned}&\tilde{b}_j(x_{j_1}, x_{j_2}, \cdots, x_{j_n}) \\ &\quad\approx \prod_{k=1}^{K} \omega_{jk}^{L_k'} \cdot \tilde{f}_{jk}\left(x_{l_1^k}, \cdots, x_{l_{L_k'}^k}\right) \\ &\quad= \prod_{k=1}^{K} \omega_{jk}^{L_k'} \cdot \prod_{d=1}^{D} \tilde{f}_{jkd}\left(x_{l_1^k d}, \cdots, x_{l_{L_k'}^k d}\right)\end{aligned} \tag{32}$$

where $\{x_{j_1}, x_{j_2}, \cdots, x_{j_n}\}$ denote feature vectors belonging to state j in $\mathbf{X}$, among which $l_1^k \cdots l_{L_k'}^k$ denote labels of the vectors "closest" to the mixture component k of state j. Then, with m_{jkd}^* and r_{jkd}^* being the pretrained mean and precision parameters, respectively, we have

$$\begin{aligned}&\tilde{f}_{jkd}(x_{1d}, x_{2d}, \cdots, x_{\zeta d}) \\ &\quad= \left(\frac{r_{jkd}^*}{2\pi}\right)^{(\zeta/2)} \cdot \frac{1}{2Cd^{-1}\rho^d} \cdot \int_{m_{jkd}^* - Cd^{-1}\rho^d}^{m_{jkd}^* + Cd^{-1}\rho^d} \\ &\qquad\cdot \exp\left(-\tfrac{1}{2}\, r_{jkd}^* \left[\sum_{t=1}^{\zeta} (x_{td} - m_{jkd})^2\right]\right) dm_{jkd} \\ &\quad= \left(\frac{r_{jkd}^*}{2\pi}\right)^{((\zeta-1)/2)} \cdot \left(\frac{1}{\zeta}\right)^{(1/2)} \cdot \frac{\Psi}{2Cd^{-1}\rho^d} \\ &\qquad\cdot \left\{\Phi\left(\sqrt{\zeta r_{jkd}^*}\left(m_{jkd}^* - \overline{x}_{\zeta d} + Cd^{-1}\rho^d\right)\right)\right. \\ &\qquad\left.-\Phi\left(\sqrt{\zeta r_{jkd}^*}\left(m_{jkd}^* - \overline{x}_{\zeta d} - Cd^{-1}\rho^d\right)\right)\right\}\end{aligned} \tag{33}$$

where $\Phi(\cdot)$ is defined in (10), and

$$\Psi = \exp\left\{-\tfrac{1}{2}\,\zeta r_{jkd}^* \left[\overline{x_{\zeta d}^2} - (\overline{x}_{\zeta d})^2\right]\right\} \tag{34}$$

with

$$\overline{x_{\zeta d}^2} = \frac{1}{\zeta} \sum_{t=1}^{\zeta} x_{td}^2$$

and

$$\overline{x}_{\zeta d} = \frac{1}{\zeta} \sum_{t=1}^{\zeta} x_{td}.$$

V. Experiments and Discussions

A simple special case of mismatch situation is encountered when the testing signal is corrupted by various additive noises, while the training data are clean. In order to examine the viability of the proposed BP-MC and VBPC algorithms, they are applied to perform speaker-independent (SI) recognition of isolated and connected digits in two sets of noisy speech recognition experiments. In the first set of experiments, the unknown mismatch is caused by additive Gaussian white noise on the testing data. While SI training is performed on clean speech data, in the testing phase, computer-generated Gaussian white noise, with various levels of intensity, is added to the original speech waveform prior to the preprocessing [25]. We also study the influence of the uncertainty neighborhood on recognition performance and report the corresponding experimental results along with our findings. In the second set of experiments, we apply the VBPC and BP-MC approaches to noisy speech recognition where 25 types of additive noises recorded in actual environments are involved. Besides, some discussions on experimental results are given to explain what mismatch situations VBPC and BP-MC with less-informative prior pdf work well and how they improve performance in these cases. In the above experiments, the degree of mismatch is measured by signal-to-noise ratio (SNR) level (dB) of the contaminated speech, which is calculated on the average over the whole testing set. No knowledge of the related mismatch is explicitly employed in testing phase. Moreover, viability of the proposed approaches on more general mismatches, e.g., mismatch caused by gender difference, is also examined. In other words, the speech models are trained on male (or female) speakers' data and then tested on female (or male) speakers' data. Finally, we also compare BP-MC and VBPC with other robust methods, including stochastic matching, minimax, and QBPC, under the mismatch caused by additive Gaussian white noise to help the readers gain some insight into the behavior of the proposed methods. In all experiments, we do not perform cepstral mean normalization in either training or testing phase.

In our recognition experiments, two speech corpora are used. The first one is called ATR Japanese isolated digits database (ATR-JPD hereafter), which is selected from ATR Japanese speech database and contains isolated utterances of Japanese 0–9 digits from 60 speakers (half male, half female).

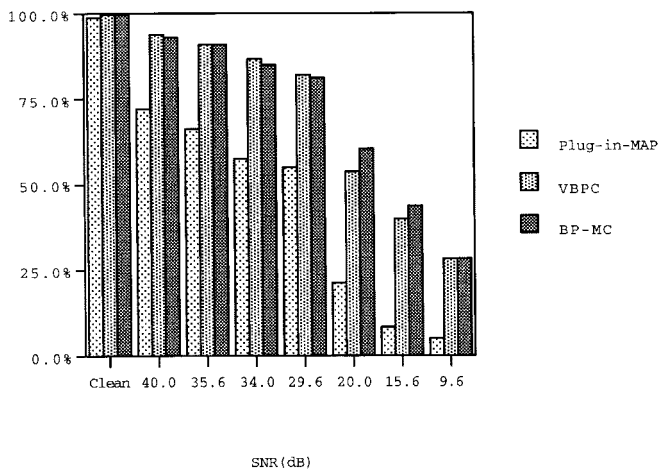


Fig. 1. Performance (word accuracy in %) comparison of VBPC and BP-MC with plug-in-MAP method at various SNR on ATR-JPD corpus when test data are distorted by Gaussian white noise.

The database ATR-JPD is recorded in a quiet environment at a sampling rate of 20 kHz with 16-b quantization accuracy. The second one is TIDIGITS English connected digit-string database [24], which includes utterances from a total of 326 speakers.

A. Noisy Speech Recognition—I: Gaussian White Noise

1) Isolated Digit Recognition: The database ATR-JPD is selected in this experiment. Each digit is modeled by a left-to-right four-state CDHMM without state skipping and each state has six Gaussian mixture components with diagonal covariance matrices. Each feature vector consists of 16 LPC-derived cepstral coefficients. For each digit, in total, we have 56 tokens from 46 speakers for speaker-independent (SI) training, and 24 tokens from other 14 different speakers for SI testing.

Fig. 1 compares the averaged recognition accuracy of the VBPC and BP-MC algorithms with that of the standard plug-in MAP based Viterbi algorithm at various SNR levels. The corresponding optimal neighborhood parameters (C, ρ) are also listed in Table I as a reference. The experimental results show that both VBPC and BP-MC are generally achieving more than 20% recognition rate improvement over that of the conventional plug-in MAP decoding in various mismatched cases. We also note that in the particular experiments here, a slight improvement is achieved even in matched condition (tested on clean speech). This suggests that the proposed techniques could also compensate for, in this case, the inaccurate estimation of model parameters caused possibly by incorrect model assumption, insufficient training data, etc.

Furthermore, we have also examined the influence of different choices of uncertainty neighborhood, i.e., neighborhood parameters C and ρ, on the final recognition performance. A similar behavior as in the minimax approach [25] that the recognition performance tends to be relatively insensitive to the shape of uncertainty regions and the performance holds up well under a wide range of SNR values is also observed in both VBPC and BP-MC. As an example, we list the recognition performance as a function of neighborhood parameters C and ρ for VBPC and BP-MC at SNR $=$ 29.6 dB in Tables II and III, respectively. Strictly speaking, the performance of VBPC and BP-MC depends on the appropriate choice of C and ρ, which in turn depends on the unknown amount of mismatch. However, the results in Tables II and III show that considerable improvement (though not optimal) can be obtained in a fairly large range of design parameters (C, ρ), which suggest that exact knowledge of C and ρ is not crucial.

2) Connected Digit Recognition: VBPC possesses the intrinsic nature of recursive search, thus VBPC can easily be extended to continuous speech recognition, with the increased cost of computation and/or memory requirement. As an example, BP-MC and VBPC are examined on TIDIGITS corpus to perform speaker-independent connected digit recognition. The feature vector consists of 12 LPC-derived cepstral coefficients, energy, and their delta features. When we are using the delta features, the mean vector m_{ik} consists of static feature in the low dimensions and delta feature in the high dimensions. The uncertainty neighborhood of Λ defined in (8) will be slightly modified as follows:

$$\begin{aligned}\eta(\Lambda) = \{\Lambda | & \pi_i = \pi_i^*,\ a_{ij} = a_{ij}^*,\ \omega_{ik} = \omega_{ik}^*,\ r_{ik} = r_{ik}^*, \\ & |m_{ikd} - m_{ikd}^*| \le C\, d^{-1} \rho^d, \\ & |m_{ik(D/2+d)} - m_{ik(D/2+d)}^*| \\ & \le C\, d^{-1} \rho^d,\ 1 \le i \le N,\ 1 \le k \le K,\ 1 \le d \le D/2\}\end{aligned} \tag{35}$$

where for $1 \le d \le D/2$, m_{ikd}'s correspond to the static feature part while $m_{ik(D/2+d)}$'s correspond to the delta feature part. The SI model for each digit is a ten-state, ten-mixture-per-state CDHMM. These digit HMM's are trained on 8623 utterances from adult training data subset of TIDIGITS. The algorithms are evaluated on 8700 utterances from the adult testing data subset distorted by various levels of computer-generated Gaussian white noises.

The recognition results of VBPC and BP-MC on TIDIGITS at several SNR levels are listed in Table IV, where **Str** stands for *string correct rate*, **Wd-C** for *word correct rate*, **Wd-A** for *word accuracy*, **Del**, **Sub**, and **Ins** for *deletion*, *substitution*, and *insertion* error rates, respectively.[3] The experimental results show that by using VBPC and BP-MC algorithms, overall recognition performance, say, digit correct rate, is improved more than 20% over that of normal plug-in-MAP based Viterbi decoding in mismatched testing conditions (SNR $=$ 36.8, 27.3, and 16.8 dB). On the other hand, VBPC and BP-MC algorithms also achieve very similar recognition performance as normal plug-in-MAP based Viterbi algorithm in matched testing condition (SNR $= \infty$) but the optimal choice of the neighborhood parameters differs from that of the mismatched case. In either mismatched or matched case, it is also observed that the recognition performance is not sensitive to different choices of neighborhood parameters in a certain region (similar to those listed in Tables II and III).

[3] All of these recognition statistics are computed by using HTK.

TABLE I
OPTIMAL NEIGHBORHOOD PARAMETERS (C, ρ) OF VBPC AND BP-MC IN FIG. 1

SNR (dB)	∞ (clean)	40.0	35.6	34.0	29.6	20.0	15.6	9.6
VBPC	(2,0.1)	(2,0.9)	(3,0.9)	(3,0.9)	(9,0.7)	(6,0.8)	(4,0.4)	(3,0.5)
BP-MC	(2,0.1)	(2,0.9)	(3,0.9)	(3,0.9)	(6,0.8)	(3,0.8)	(3,0.7)	(4,0.7)

TABLE II
RECOGNITION ACCURACY (IN %) AS A FUNCTION OF NEIGHBORHOOD PARAMETERS C AND ρ OF VBPC AT SNR = 29.6 dB (PLUG-IN-MAP ATTAINS 55% CORRECT RATE)

C \ ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	54.2	54.6	54.6	53.3	54.6	55.4	54.6	60.4	66.3
2	54.6	53.3	52.5	56.7	62.9	67.9	65.8	70.4	69.6
3	53.8	52.9	60.0	65.4	65.0	70.0	69.2	71.3	77.9
4	53.8	55.0	64.2	67.9	70.4	70.0	67.9	76.7	71.7
5	53.3	62.1	62.1	70.0	68.8	64.6	73.3	80.0	50.8
6	52.9	62.9	69.2	68.3	66.3	67.5	76.7	78.8	36.3
7	52.9	62.9	70.8	67.9	63.3	67.9	78.3	78.3	27.1
8	55.8	64.6	67.5	62.1	63.3	70.0	79.6	75.4	25.8
9	59.6	67.1	63.8	63.8	64.2	70.8	82.1	72.9	23.8
10	62.1	70.0	62.9	63.3	67.5	74.6	80.8	67.9	20.8
11	63.8	70.4	64.6	62.1	67.9	74.2	81.7	62.9	18.3
12	62.5	69.2	64.6	64.2	70.0	75.8	80.8	58.8	16.3
13	64.2	66.7	64.6	62.5	69.6	77.1	80.8	54.2	12.9
14	63.3	64.6	63.8	65.8	70.4	73.8	79.2	48.3	13.3
15	63.3	62.1	64.6	64.2	69.6	75.8	77.1	39.6	14.2
16	63.3	61.7	62.9	65.4	68.8	75.8	77.1	36.3	12.1
17	65.0	61.3	64.6	64.2	70.0	77.9	75.0	34.6	12.1
18	66.3	61.3	63.3	64.2	67.9	77.9	75.0	34.6	10.0
19	67.9	61.3	62.1	63.3	71.3	76.7	70.0	35.8	10.4
20	68.3	60.0	61.7	64.6	69.2	78.3	71.7	30.8	12.1

TABLE III
RECOGNITION ACCURACY (IN %) AS A FUNCTION OF NEIGHBORHOOD PARAMETERS C AND ρ OF BP-MC AT SNR = 29.6 dB (PLUG-IN-MAP ATTAINS 55% CORRECT RATE)

C \ ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	56.3	56.3	57.5	58.3	58.8	62.1	64.2	65.8	67.1
2	57.1	57.9	60.8	64.6	68.3	68.8	70.0	74.6	75.4
3	57.5	60.0	67.1	68.3	71.7	72.1	72.5	74.6	77.9
4	58.8	63.8	67.9	70.0	69.2	68.8	70.0	77.1	76.3
5	58.3	67.1	68.8	70.0	67.1	67.1	70.0	79.2	66.3
6	59.6	67.1	68.3	67.5	63.8	67.1	73.3	81.3	55.8
7	61.7	68.3	66.7	64.2	62.5	65.4	75.4	81.3	47.1
8	63.8	70.4	64.6	61.3	63.3	67.1	76.3	80.4	41.3
9	65.4	70.8	63.8	61.7	63.8	69.2	76.3	78.3	33.3
10	67.9	67.9	63.3	62.9	65.4	71.3	79.6	77.5	31.3
11	67.9	67.1	61.7	62.9	66.7	71.7	80.0	75.0	28.3
12	67.1	65.8	61.3	63.3	65.4	72.9	80.4	70.8	23.3
13	67.1	65.4	61.7	62.9	65.4	72.1	78.8	65.8	20.8
14	67.5	66.3	61.3	62.9	64.6	71.3	77.9	63.8	20.8
15	68.8	63.8	61.3	63.3	65.0	71.3	78.3	62.5	20.4
16	69.6	62.5	62.5	64.2	66.3	71.7	77.9	59.6	19.2
17	69.2	61.7	62.1	63.8	66.3	71.3	78.3	56.3	16.7
18	67.9	60.8	62.1	63.3	66.3	72.5	79.6	55.0	14.6
19	67.9	61.3	62.5	64.2	67.1	74.2	78.8	52.5	12.5
20	67.5	60.8	63.8	64.2	67.9	75.4	78.8	50.8	13.8

B. Noisy Speech Recognition—II: Real-World Noises

An attempt has also been made to cover a class of mismatch situations as wide and general as possible. In this section, we evaluate the proposed algorithms in noisy speech recognition where 25 types of additive noises recorded in actual environments [14] are involved.

1) Description of Actual Noises: We choose the Japan Electronic Industry Development Association (JEIDA) noise database for our experiments. The included noise data are collected in various kinds of environments under which speech input devices are expected to be typically used [14]. Their characteristics are summarized in Table V. From Table V, we notice that these noises differ much in both nature and characteristics. Most of these noises are very difficult to deal with because they are nonstationary in time domain, and have a complex spectrum with a wide bandwidth.

2) Noisy Speech Recognition Results: We first evaluate VBPC on noisy speech recognition problem involving the above mentioned 25 types of actual noises. Our task is again isolated digit recognition on corpus ATR-JPD. The experimental setup is the same as that of Section V-A1. The mismatch between test and training conditions is caused by adding those actual noises on test data at various SNR levels. It is not easy to define a proper SNR measure for nonstationary signals [1]. In this study, we simply adopt an SNR measure, which is defined as the ratio between the signal variance and the noise variance. This SNR measure only reflects the general size or degree of mismatches caused by adding the related noises. Various noises are scaled to achieve several SNR levels before added to the clean speech.

TABLE IV
PERFORMANCE (IN %) COMPARISON OF VBPC AND BP-MC WITH PLUG-IN-MAP METHOD ALGORITHM ON TIDIGITS CORPUS WHEN TEST DATA ARE DISTORTED BY GAUSSIAN WHITE NOISE

SNR		Str	Wd-C	Wd-A	Del	Sub	Ins
∞	Plug-in-MAP	89.95	98.91	97.76	0.31	0.79	1.15
	BP-MC	87.93	98.42	97.27	0.68	0.90	1.14
	VBPC	89.51	98.70	97.60	0.37	0.94	1.09
36.8(dB)	Plug-in-MAP	17.83	67.49	66.39	16.11	16.40	1.11
	BP-MC	62.55	91.35	89.42	3.98	4.66	1.94
	VBPC	61.08	90.30	89.09	5.13	4.57	1.21
27.3(dB)	Plug-in-MAP	0.20	45.19	43.80	25.15	29.66	1.39
	BP-MC	39.42	80.58	78.56	10.47	8.96	2.01
	VBPC	28.95	73.97	73.13	15.38	10.65	0.84
16.8(dB)	Plug-in-MAP	0.0	25.05	24.04	45.24	29.71	1.01
	BP-MC	11.87	56.07	54.98	29.39	14.54	1.09
	VBPC	5.91	45.86	45.47	38.41	15.74	0.39

We depict the experimental results of VBPC and plug-in-MAP method at SNR levels 0, 10, 20, and 30 dB on the above 25 actual noises, respectively, in Figs. 2–4.[4] The experimental results clearly show that VBPC approach works well for most of these actual noises under a wide range of SNR values (we call these cases Type I, such as noises no. 6, 11, 16, 21, 22, 24,

[4]The results on noise no. 17 are not included here because we had data-error in reading noise no. 17 from disc.

TABLE V
TWENTY-FIVE TYPES OF ACTUAL NOISES USED IN THE EXPERIMENTS

No.	Noise description	Bandwidth	Spectrum	Stationary
1	Automobile cabin in street(Medium-size car)	Narrow	simple,smooth	○
2	Automobile cabin in highway(Medium-size car)	Narrow	simple,smooth	○
3	Automobile cabin in highway(Compact car)	Narrow	simple,smooth	○
4	Automobile cabin in street(Compact car)	Narrow	simple,smooth	○
5	Exhibition hall A(In a booth)	wide	complex,jagged	×
6	Exhibition hall B(In a passage)	Middle	simple,smooth	×
7	Railway station(Near ticket vending machines)	Wide	complex,jagged	×
8	Railway station (In a passage)	Middle	complex,jagged	×
9	Telephone booth (Downtown)	Narrow	simple,smooth	○
10	Factory (Machinery)	Wide	complex,jagged	×
11	Factory (Press)	Wide	complex,jagged	×
12	Parcel classification works (1)	Wide	complex,jagged	×
13	Parcel classification works (2)	Middle	complex,jagged	×
14	Trunk road	Wide	complex,jagged	×
15	Road crossing	Middle	one high-freq comp.	×
16	Crowded street	Middle	smooth	×
17	New trunkline train	Narrow	complex,jagged	×
18	Ordinary train	Wide	simple,smooth	×
19	Computer room A (Minicomputers)	Wide	complex,jagged	○
20	Computer room B (Workstations)	Middle	smooth	○
21	Large air conditioner	Middle	jagged	○
22	Air conditioning fan coil	Middle	jagged	○
23	Ventilation duct	Wide	jagged	×
24	Elevator passage (Hospital)	Narrow	smooth	×
25	Elevator passage (Department store)	Middle	complex,jagged	×

etc.) and is also helpful for the remaining ones (we call these cases Type II, such as noises no. 15, 19, and 25). It is quite encouraging that the VBPC is effective for a great variety of mismatches examined here.

Once again, we observe that the performance of VBPC is fairly insensitive to the hyperparameters C and ρ in these experiments. To show this, we list in Table VI, the optimal recognition rates of VBPC averaged over four SNR levels (0, 10, 20, and 30 dB) in each of above 25 noise types, as well as the corresponding optimal values of (C, ρ). For comparison, the corresponding results of conventional plug-in-MAP method and that of VBPC at $(C = 2, \rho = 0.9)$ are also listed. For type I mismatch, VBPC works well by choosing a relatively wider neighborhood, i.e., $C\epsilon[1, 3]$, $\rho\epsilon[0.8, 0.9]$. For type II mismatch, VBPC only works by choosing a relatively smaller size of the neighborhood, i.e., $C = 1$, $\rho\epsilon[0.1, 0.3]$. However, no major performance improvement is observed in type II case. It is expected that the performance of VBPC will converge to that of the normal plug-in MAP when the size of the neighborhood approaches to zero. In the specific experiments here, the results in Table VI suggest that $(C = 2, \rho = 0.9)$ is an acceptable choice for most noises.

We have also examined the BP-MC approach in these 25 types of noises and a similar behavior as the VBPC is observed. As an example, the results of five types are shown in Fig. 5. From the results in Figs. 2–5, we notice that one *necessary* condition for VBPC and BP-MC approaches to profit most is that the size and/or degree of mismatch is comparable with the "distance" between models in some sense. The benefit of using VBPC and BP-MC approaches decreases as the size and/or degree of mismatches become too small or too large. The unconfusable vocabulary task, which implies large enough "distance" between models, warrants that the VBPC and BP-MC approaches have more chances to work well for various mismatches with different nature and degree. Although the discussion here is based on isolated word recognition results, the same behavior can be observed and the same conclusion can also be drawn from experimental results on connected word recognition.

C. Cross-Gender Speech Recognition

We have also examined the viability of the proposed algorithms in a more general mismatch caused by gender difference. The corpus ATR-JPD is chosen again for the cross-gender recognition experiment. The gender-dependent models are trained by male (or female) speech data but tested on female (or male) speech data. The experimental results are shown in Table VII. The recognition results show that VBPC and BP-MC also work in the case of cross-gender mismatch, but improvement is generally minor. Only about 3–5% absolute recognition rate improvement is achieved on the average.

D. Comparative Study with Other Robust Methods

In this section, we present a comparative study of BP-MC and VBPC with other robust techniques, including QBPC [12],

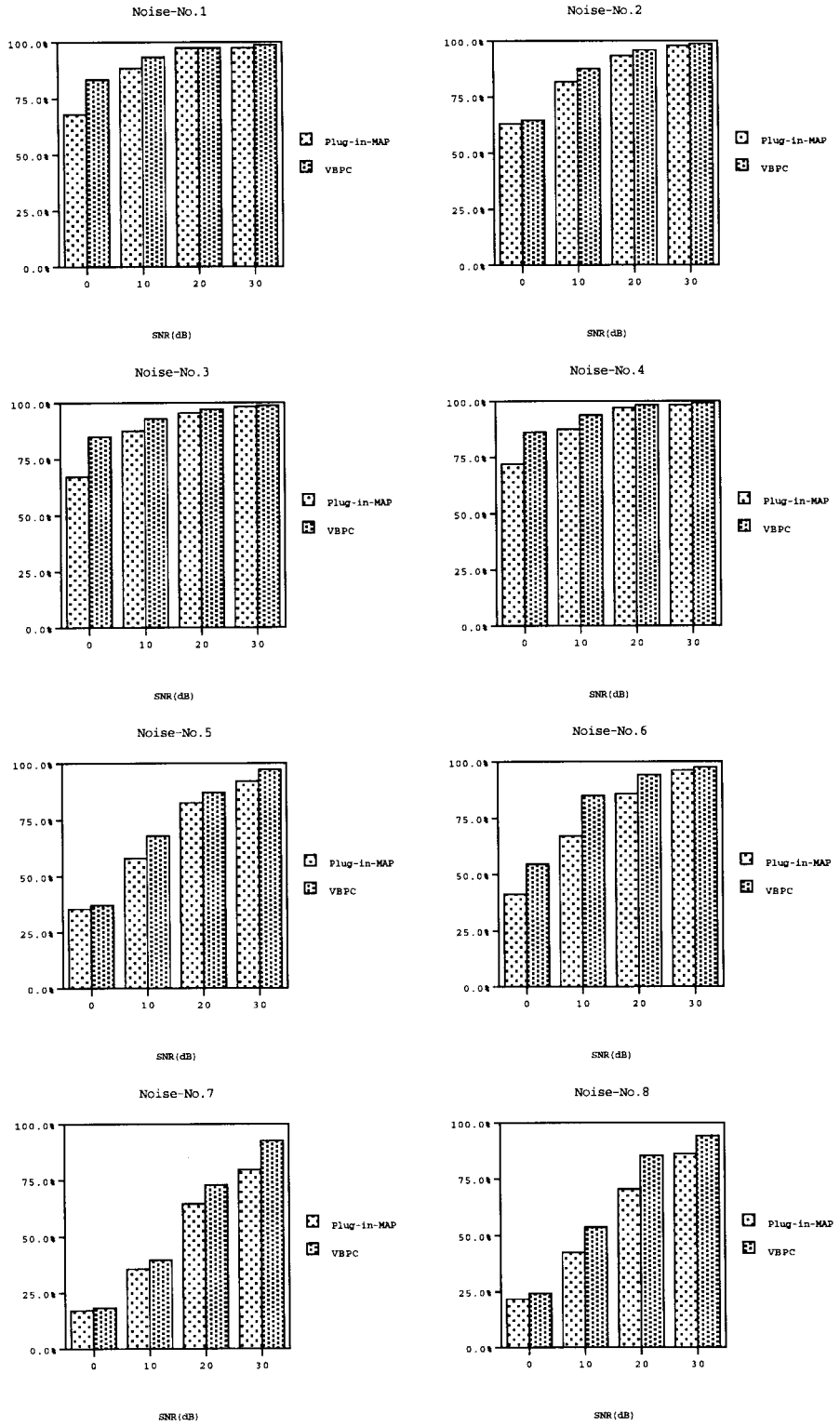


Fig. 2. Performance (word accuracy in %) comparison of VBPC with normal plug-in-MAP method on 25 types of actual noises at SNR = 0, 10, 20, 30 (dB): Part I.

[13], minimax [25], and stochastic matching (SM) [29], under the condition that neither the knowledge of mismatches nor adaptation data is available. The experimental setup is the same as in Section V-A1. The mismatch between training and testing conditions is caused by adding Gaussian white noise into test data prior to the preprocessing at three SNR levels (10, 20, and 30 dB's, respectively).

The detailed description about QBPC can be found in [12] and [13]. In the current experiment, for simplicity, we only implement the Viterbi version of the quasi-Bayes approximation in QBPC computation. We also only consider the uncertainty of the mean vectors and only one iteration is performed at each QB approximation step. The prior pdf is chosen as the best normal approximation to the constrained uniform distribution

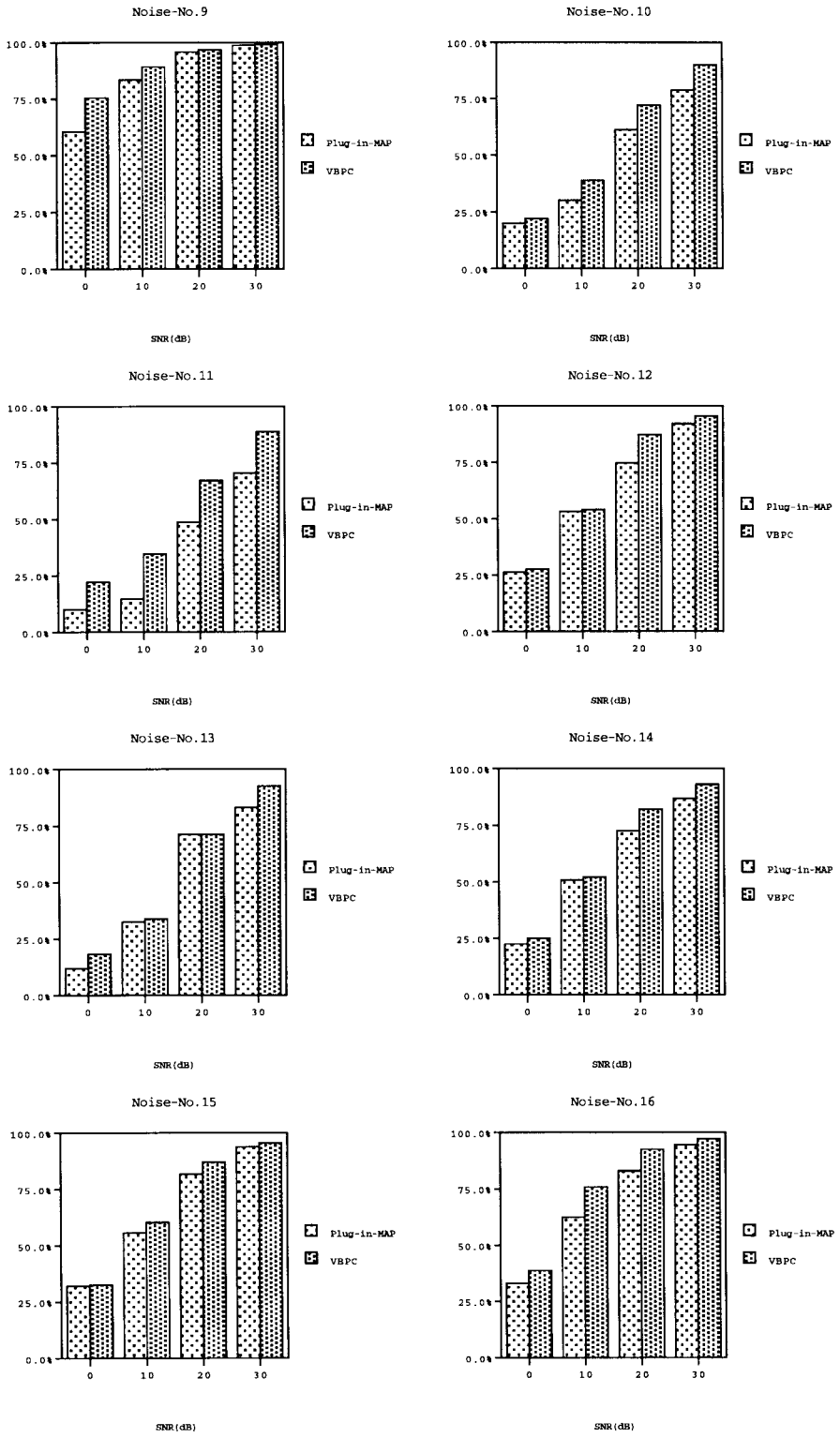


Fig. 3. Performance (word accuracy in %) comparison of VBPC with normal plug-in-MAP method on 25 types of actual noises at SNR = 0, 10, 20, 30 (dB): Part II (continued).

(8) to minimize the Kullback–Leibler directed divergence (see [13] for the details).

As for stochastic matching, as discussed in [29], we compensate for the mismatch in either feature space or model space. In the feature space method (denoted as SM-FS1), a single fixed additive bias in cepstral domain is used for each utterance. In the model space method (denoted as SM-MS1), a single random bias with a Gaussian pdf is adopted. In both methods, the bias vector or the mean of the random bias is initialized to zero. In model space method, the variance of the random bias is initialized to a small positive number. Two to five iterations are performed for the ML *nuisance parameters* estimation.

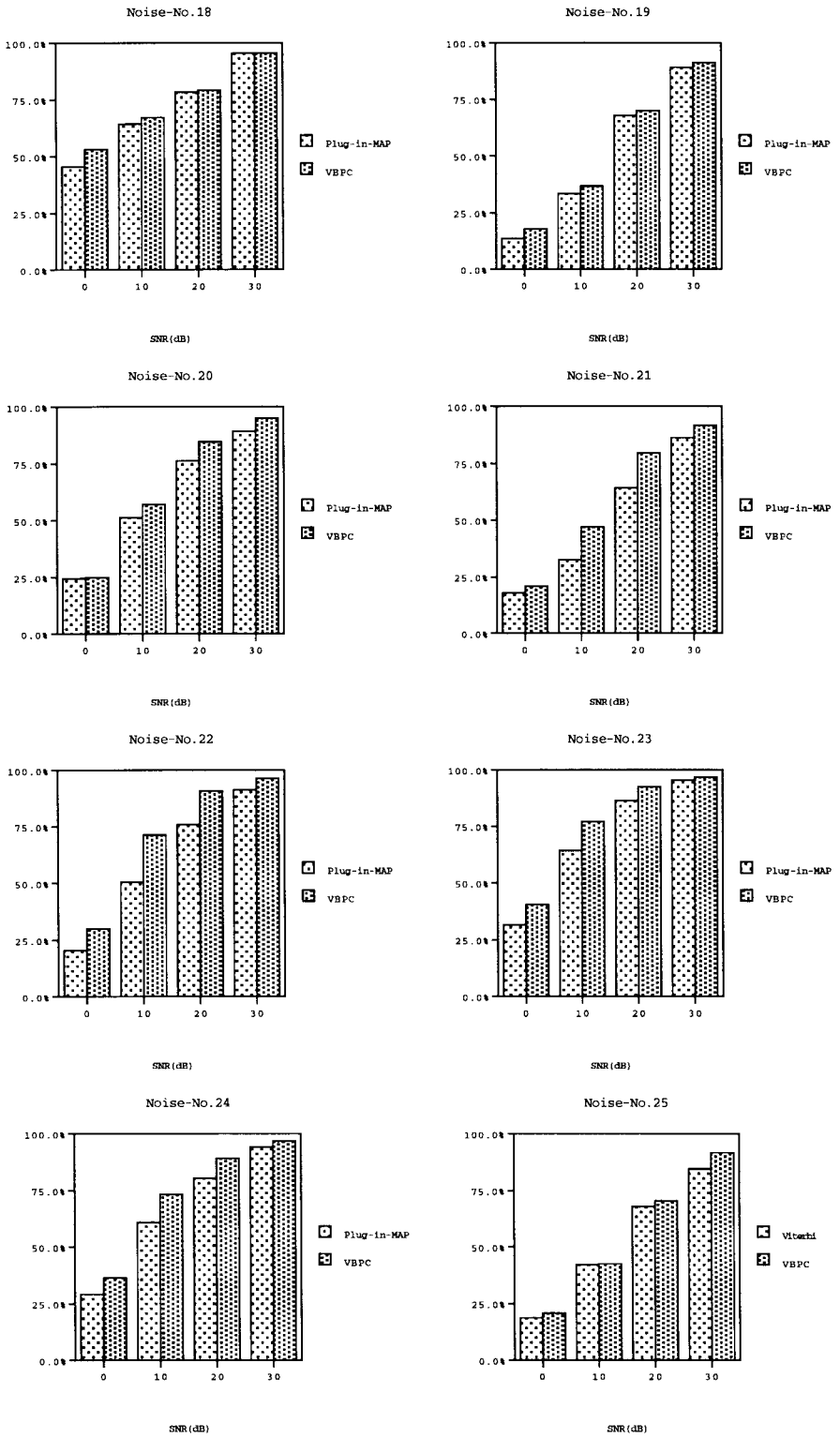


Fig. 4. Performance (word accuracy in %) comparison of VBPC with normal plug-in-MAP method on 25 types of actual noises at SNR = 0, 10, 20, 30 (dB): Part III (continued).

In [25], Merhav and Lee perform the minimax classification as in (2), where the parameter neighborhood Ω is assumed to follow (8). In their implementation, to approximate $\max_{\Lambda \in \Omega} p(\mathbf{X}|\Lambda, W)$ in (2), the following iterative procedure is used.

- Initialize Λ with the values obtained in the training phase.
- In each iteration, first decode the optimal path s^*, l^* using the Viterbi algorithm; then the model parameter Λ is reestimated according to s^*, l^*.
- If the new Λ falls in Ω, it is used to update the old Λ; otherwise, the parameter within Ω which is closest to the new Λ is chosen.

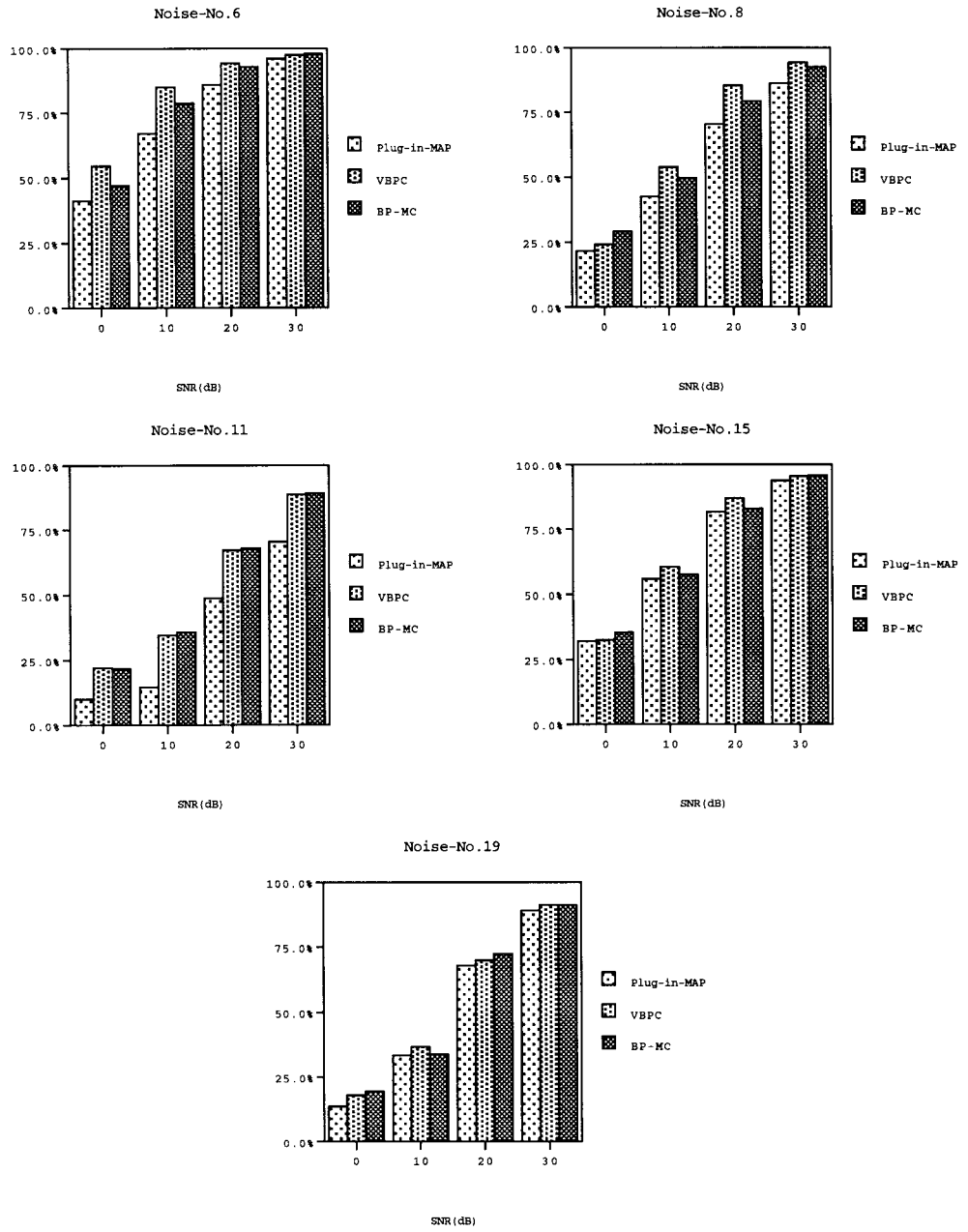


Fig. 5. Performance (word accuracy in %) comparison of BP-MC, VBPC with normal plug-in-MAP method on five selected types of actual noises at SNR = 0, 10, 20, 30 (dB).

In this paper, Merhav and Lee's minimax is denoted as minimax1. Besides, another so-called modified minimax used in [13] works as follows:

$$\hat{W} = \arg\max_{W} p(\mathbf{X}|\Lambda_{\mathrm{MAP}}, W) \tag{36}$$

where

$$\Lambda_{\mathrm{MAP}} = \arg\max p(\mathbf{X}|\Lambda, W) \cdot p(\Lambda|\varphi, W)$$

with the prior pdf $p(\Lambda|\varphi, W)$ chosen in the same way as the QBPC. This modified minimax method is denoted as minimax2 here.

The experimental results of these methods are compared in Table VIII. In the table, for VBPC, BP-MC, QBPC, minimax1, and minimax2, we only show the best performance achieved under the optimal choice of hyperparameters (i.e., C and ρ) within a certain range (i.e., $C\epsilon[1, 10]$ and $\rho\epsilon[0.1, 0.9]$). According to the results, several observations can be made. At first, the results show that both BP-MC and VBPC outperform the Viterbi implementation of the QBPC. It is experimentally shown here that the VBPC achieves a better approximation of the true BPC than QBPC, but at the expense of much higher computational overhead. Second, as expected, we note that the performance improvement of the stochastic matching is moderate, especially in the low SNR level. However, several points should be noted in this comparative experiment. In

TABLE VI
RECOGNITION RATES OF NORMAL PLUG-IN-MAP METHOD AVERAGED OVER FOUR SNR LEVELS OF (0, 10, 20, AND 30 dB) IN EACH OF 25 NOISE TYPES, CORRESPONDING OPTIMAL RECOGNITION RATES OF VBPC WITH OPTIMAL VALUES OF (C, ρ), RECOGNITION RATES OF VBPC WITH $(C = 2, \rho = 0.9)$

Noise No.	Plug-in-MAP	Optimal VBPC	Optimal (C, ρ)	VBPC at (C=2,ρ=0.9)
1	87.81%	91.67%	(2,0.9)	-
2	83.96%	85.21%	(2,0.9)	-
3	87.08%	92.40%	(2,0.9)	-
4	88.75%	93.23%	(2,0.9)	-
5	67.08%	70.62%	(2,0.9)	-
6	72.81%	82.08%	(2,0.9)	-
7	49.49%	55.83%	(2,0.9)	-
8	55.21%	61.77%	(2,0.9)	-
9	84.58%	88.54%	(2,0.9)	-
10	47.50%	52.71%	(1,0.9)	50.73%
11	35.94%	48.85%	(2,0.9)	-
12	61.56%	62.40%	(2,0.9)	-
13	49.69%	51.56%	(1,0.9)	49.72%
14	58.12%	61.04%	(1,0.9)	59.48%
16	68.23%	75.31%	(2,0.9)	-
18	70.83%	72.81%	(1,0.9)	70.83%
20	60.31%	64.48%	(2,0.9)	-
21	50.21%	57.81%	(3,0.9)	56.77%
22	59.48%	70.10%	(2,0.9)	-
23	69.38%	74.27%	(2,0.9)	-
24	66.15%	73.23%	(2,0.9)	-
15	65.83%	65.83%	(1,0.1)	65.73%
19	51.04%	51.24%	(1,0.2)	47.29%
25	53.33%	53.85%	(1,0.3)	51.15%

the mismatch situation caused by additive noise, the bias is additive in linear spectral domain, but not in the cepstral domain. Therefore, the bias compensation in cepstral domain makes the assumption behind the ML-based stochastic matching method invalid. But when we adopt cepstrum based feature, the ML-based stochastic matching in linear spectral domain does not possess a straightforward form to implement. Moreover, ML-based stochastic matching approach is a fully automatic procedure where ML criterion helps to determine all the nuisance parameters. In all of the other robust methods we studied here, including VBPC, BP-MC, QBPC, minimax1, and minimax2, we have to manually choose a few control parameters (i.e., C and ρ). The optimal performance of these methods shown in the Table VIII are better than that of the stochastic matching. We also observe in the experiments that, apart from the optimal choice of C and ρ, these methods also perform better than the SM in a certain range of C and ρ. Next, the BPC performance depends heavily on the appropriate choice of the prior distribution. In the case of the mismatch caused by the additive Gaussian white noise, the chosen prior distribution seems to be able to model the mismatch appropriately. At last, we also notice that both the VBPC and the BP-MC outperform minimax1 and minimax2 in this case.

E. Discussions

In principle, the methodology of BP-MC and VBPC is suitable to any possible mismatches. By using a less-informative prior pdf in this study, the algorithms are quite flexible that we only need to adjust/adapt *two* hyperparameters C and ρ to deal with a great variety of mismatches. However, strictly speaking, the performance of these methods depends on the appropriate choice of neighborhood (e.g., the values of C and ρ), which in turn depends on the unknown amount and nature of the mismatch. Although it has been experimentally shown that the performance is not sensitive to the choice of neighborhood in the examined mismatch conditions, as mentioned above, if we expect to benefit most from the method, e.g., to deal with simultaneously various types of possible mismatches in practice, it will be important to develop a simple online adjusting procedure to tune the neighborhood parameters based on only very few training/adaptation data for attaining the optimal performance in various cases (e.g., both mismatched and matched cases). This remains a topic for future research.

TABLE VII
PERFORMANCE (WORD ACCURACY IN %) COMPARISON OF VBPC, BP-MC WITH PLUG-IN-MAP METHOD ON CROSS-GENDER ISOLATED DIGIT RECOGNITION TASK

Training data	Testing data	Plug-in-MAP	VBPC	BP-MC
SI	SI	98.8	99.6	99.6
Male	Female	73.6	76.1	75.3
Female	Male	49.0	53.8	54.0

In terms of computational complexity, BP-MC is obviously more costly than the conventional plug-in-MAP based Viterbi decoding in computing component densities [e.g., (9)], in either isolated word or continuous speech recognition. Since these calculations usually are followed by a log operation (or a table look-up), the increased cost is not negligible, especially in large vocabulary applications. Like BP-MC, VBPC also consumes many more computations in calculating

TABLE VIII
PERFORMANCE (WORD ACCURACY IN %) COMPARISON OF VBPC, BP-MC WITH PLUG-IN-MAP, QBPC, STOCHASTIC MATCHING (SM-FS1 AND SM-MS1), MINIMAX (MINIMAX1), AND MODIFIED MINIMAX (MINIMAX2) WHEN TEST DATA ARE DISTORTED BY GAUSSIAN WHITE NOISE. (THE NUMBERS IN PARENTHESES DENOTE THE OPTIMAL NEIGHBORHOOD PARAMETERS (C,ρ) FOR THE CORRESPONDING METHOD TO ACHIEVE THE SHOWN PERFORMANCE AT EACH CASE)

SNR	Plug-in-MAP	SM-FS1	SM-MS1	VBPC	BP-MC	QBPC	minimax1	minimax2
30(dB)	62.08	65.0	68.75	79.17 (6,0.7)	82.81 (10,0.7)	71.25 (1,0.7)	73.33 (2,0.5)	71.67 (1,0.4)
20(dB)	26.10	31.25	32.92	60.83 (6,0.3)	62.92 (9,0.4)	45.83 (1,0.7)	57.92 (3,0.4)	53.33 (1,0.6)
10(dB)	5.42	6.25	9.58	33.33 (7,0.3)	37.08 (9,0.4)	24.58 (2,0.5)	28.33 (7,0.3)	26.25 (5,0.2)

the predictive density than the conventional algorithm [e.g., (33)]. Besides, extra efforts such as the repeated backtracing and the related bookkeeping are further required in VBPC search procedure. Although this might be affordable in small vocabulary recognition tasks, either for isolated words or continuous speech, we will encounter serious computational difficulty when we apply the VBPC method to large vocabulary continuous speech recognition. This is because during the frame-synchronous search of the VBPC, the computation of the partial predictive pdf depends on the hypothesized partial optimal path up to each time instant. In a large vocabulary case, this will easily lead to a combinatorial explosion, thus make the problem untractable. Therefore, some simplified schemes are needed to make the VBPC algorithm computationally feasible. For instance, a narrow *beam* VBPC search strategy might mitigate the difficulty somehow. We can also use the normal search algorithms to first obtain an N-best list of the possible paths and then select the final results from these N-best paths based on the VBPC approach.

VI. CONCLUSION

In this paper, we have studied a category of robust speech recognition problem from the viewpoint of Bayesian prediction. A Bayesian predictive density based model compensation (BP-MC) technique and a robust decision strategy called Viterbi Bayesian predictive classification (VBPC) are presented in this study. To examine the viability of the proposed techniques, BP-MC and VBPC are performed on speaker-independent isolated digit and connected digit string recognition tasks, where severe mismatches exist between training and testing conditions. Following are some of our findngs.

- Both BP-MC and VBPC approaches improve the performance robustness under various mismatches examined even when we have little prior knowledge about these mismatches. This suggests that Bayesian prediction could be a potential approach to achieve the robustness in speech recognition.
- The less-informative prior pdf adopted in this study is straightforward and only two uncertainty neighborhood parameters need to be tuned in advance. We experimentally show that under a wide range of values of these control parameters, the proposed techniques help in improving the performance when some mismatches exist between training and testing conditions. If the prior pdf can characterize the actual mismatches in question, we will have a larger chance to improve the performance. Otherwise, no considerable gain will be expected.

Apart from the issues we discussed in the previous sections, we are still not sure whether VBPC and BP-MC formulations can work well in a more confusable vocabulary case because these methods improve the performance robustness in mismatched conditions at the expense of decreasing the discriminative ability of the models. Moreover, although VBPC and BP-MC improve the performance over the nonrobust method in the mismatched cases we examined, the absolute recognition rate of VBPC and BP-MC in mismatched case is still far inferior to matched condition results. How to bridge this performance gap is still a challenging topic for further research.

REFERENCES

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Boston, MA: Kluwer, 1993.

[2] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 179–190, Mar. 1983.

[3] S. Furui, "Recent advances in robust speech recognition," in *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, Apr. 1997, pp. 11–20.

[4] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, 1994.

[5] S. Geisser, *Predictive Inference: An Introduction.* London: Chapman & Hall, 1993.

[6] Y.-F. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, 1995.

[7] P. J. Huber, *Robust Statistics.* New York: Wiley, 1981.

[8] Q. Huo, C. Chan, and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 334–345, 1995.

[9] ———, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 141–144, 1996.

[10] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 161–172, 1997.

[11] ———, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 386–397, July 1998.

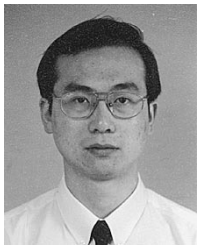
[12] Q. Huo, H. Jiang, and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," in *Proc. ICASSP'97*, Munich, Germany, Apr. 1997, vol. II, pp. 1547–1550.

[13] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," submitted for publication.

[14] S. Itahashi, "Creating speech corpora for speech science and technology," *Trans. IEICE*, vol. E74, pp. 1906–1910, 1991.

[15] H. Jiang, K. Hirose, and Q. Huo, "A CDHMM-based robust speech recognition approach using plug-in predictive density of Gaussian mixture component," in *Proc. Acoustical Soc. Jpn. Fall Meeting*, Okayama, Japan, Sept. 1996, pp. 149–150.

[16] ——, "Robust speech recognition based on Viterbi Bayesian predictive classification," in *Proc. ICASSP'97,* Munich, Germany, Apr. 1997, vol. II, pp. 1551–1554.
[17] ——, "Robust speech recognition based on Bayesian predictive approach," in *Tech. Rep. IEICE,* SP96-93 (1997-01), Jan. 1997, pp. 45–52 (in English).
[18] ——, "Applying Viterbi Bayesian predictive classification to robust recognition of continuous speech," in *Proc. Acoustical Soc. Jpn. Spring Meeting,* Kyoto, Japan, Mar. 1997, pp. 35–36.
[19] ——, "Use of less-informative Bayesian predictive classification for noisy speech recognition," in *Proc. 1997 China-Japan Symp. Advanced Information Technology,* Huangshan, China, Apr. 1997, pp. 41–46.
[20] B.-H. Juang, "Speech recognition in adverse environments," *Comput. Speech Lang.,* vol. 5, pp. 275–294, 1991.
[21] C.-H. Lee, F.-K. Soong, and K.-K. Paliwal, Eds., *Automatic Speech and Speaker Recognition: Advanced Topics.* Boston, MA: Kluwer, 1996.
[22] C.-H. Lee, "On feature and model compensation approach to robust speech recognition," in *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels,* Pont-a-Mousson, France, Apr. 1997, pp. 45–54.
[23] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.,* vol. 9, pp. 171–185, 1995.
[24] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP'84,* pp. 42.11.1–4.
[25] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition," *IEEE Trans. Speech Audio Processing,* vol. 1, pp. 90–100, 1993.
[26] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Signal Processing,* vol. 39, pp. 2157–2166, 1991.
[27] A. Nadas, "Optimal solution of a training problem in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 326–329, 1985.
[28] R. C. Rose, C.-H. Lee, and B.-H. Juang, "Model compensation for robust ASR," in *Proc. IEEE ASR Workshop,* Snowbird, UT, Dec. 1995, pp. 98–100.
[29] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Processing,* vol. 4, pp. 190–202, May 1996.
[30] B. M. Shahshahani, "A Markov random field approach to Bayesian speaker adaptation," in *Proc. ICASSP'96,* Atlanta, GA, pp. 697–700.



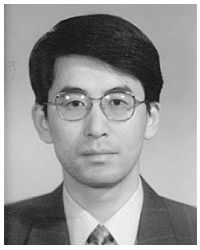
Hui Jiang was born in Kunming, China, in 1970. He received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China (USTC), Hefei, in 1992 and 1994, respectively, and the Ph.D. degree from the University of Tokyo, Tokyo, Japan, in 1998, all in electrical engineering. His Ph.D. dissertation was involved with robust speech recognition.

From 1992 to 1994, he worked on large vocabulary Chinese speech recognition at USTC. From 1998 to 1999, he worked with a Japanese national project involving research on man–machine dialogue systems through spoken languages at the University of Tokyo, where he worked on large vocabulary continuous speech recognition of the Japanese language. Since April 1999, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His current research interests include all issues related to speech recognition and understanding, especially robust speech recognition, utterance verification, adaptive modeling of speech, and dialogue system design.

Keikichi Hirose (M'78) received the B.E. degree in electrical engineering in 1972, and the M.E. and Ph.D. degrees in electronic engineering in 1974 and 1977, respectively, from the University of Tokyo, Tokyo, Japan.

In 1977, he joined the University of Tokyo as a Lecturer in the Department of Electrical Engineering. He has been a Professor at the Graduate School for Information and Communication Engineering, University of Tokyo, since April 1994. In April 1999, he received a dual appointment as Professor in the University's Graduate School of Frontier Sciences. From March 1987 to January 1988, he was a Visiting Scientist of the Research Laboratory of Electronics, Massachusetts Institute of Technology Cambridge. His research interests include prosody, speech synthesis and recognition, and CALL systems.

Dr. Hirose is a member of the Acoustical Society of America, the European Speech Communication Association, the Institute of Electronics, Information and Communication Engineers, the Acoustical Society of Japan, the Information Processing Society of Japan, and other professional organizations.



Qiang Huo (M'95) received the B.Eng. degree from the University of Science and Technology of China (USTC), Hefei, in 1987, the M.Eng. degree from Zhejiang University, Hangzhou, China, in 1989, and the Ph.D. degree from the USTC, in 1994, all in electrical engineering. He completed the Ph.D. dissertation on speech recognition under a joint training program between Hong Kong University (HKU) and USTC.

From 1986 to 1990, his research work focused on the hardware design and development for real-time digital signal processing, image processing and computer vision, and speech and speaker recognition. From 1991 to 1994, he was with the Department of Computer Science, HKU. From 1995 to 1997, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he engaged in research in speech recognition. He joined the Department of Computer Science and Information Systems, HKU again in January 1998 as an Assistant Professor. His current major research interests include speech/speaker recognition, computational model for spoken dialogue processing, Chinese character recognition, adaptive signal modeling and processing, artificial neural network algorithms, and general pattern recognition theory.